Cognitive Psychology 136 (2022) 101495

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

What is theory of mind? A psychometric study of theory of mind and intelligence

Ester Navarro

Center for Applied Brain and Cognitive Sciences, Tufts University, School of Engineering, 177 College Ave, Medford, MA 02155, United States

ARTICLE INFO

Keywords:
Social cognition
Theory of mind
Psychometric measurement
Intelligence

ABSTRACT

Theory of mind (ToM) is an essential ability for social competence and communication, and it is necessary for understanding behaviors that differ from our own (Premack & Woodruff, 1978). Recent research suggests that tasks designed to measure ToM do not adequately capture a single ToM ability (Warnell & Redcay, 2019; Quesque & Rossetti, 2020) and, instead, might be related to tasks of general cognitive ability (Coyle, Elpers, Gonzalez, Freeman, & Baggio, 2018). This hinders the interpretation of experimental findings and puts into question the validity of the ToM construct. The current study is the first psychometric assessment of the structure of ToM to date. Comparing ToM to crystallized intelligence (Gc) and fluid intelligence (Gf), the study aims to (a) understand whether ToM should be considered a monolithic ability and (b) explore whether tasks of ToM adequately assess ToM, above and beyond general cognitive ability. For this, confirmatory factor analyses (CFAs), exploratory factor analysis (EFA), and exploratory network analysis (NMA) were conducted. The results of the models largely point to the same conclusion: while ToM tasks are not merely assessing cognitive ability, they are not purely assessing a single ToM construct either. Importantly, these findings align with recent theoretical accounts proposing that ToM should not be considered a monolithic construct (Quesque & Rossetti, 2020; Schaafsma, Pfaff, Spunt, & Adolphs, 2015, Devaine, Hollard, & Daunizeau, 2014), and should instead be explored and measured as multiple domains.

1. Introduction

How do humans understand what other people feel and believe? Psychologists and philosophers have long asked this basic question. Theory of mind (ToM) is considered the ability to understand the beliefs, knowledge, and intentions of others based on their behavior. The term was first coined by Premack and Woodruff (1978) to refer to chimpanzees' ability to infer human goals, and it was quickly adopted by psychologists to study humans' ability to infer and predict the behavior of others. This was followed by a vast number of studies on the topic. A simple search of the term "theory of mind" on PsycInfo (2020) reveals over 7000 articles and 1000 books with Theory of Mind on the title. Surprisingly, while the relationship between ToM and a multiplicity of constructs has been thoroughly studied (e.g., communication, Grice, 1989; Sperber & Wilson, 1995: understanding criticism, Cutting & Dunn, 2002; deception, Sodian, 1991; joking and lying, Hughes & Leekam, 2004; Leekam & Prior, 1994; irony, Happé, 1994; pragmatic language competence, Eisenmajer & Prior, 1991; problem solving, Greenberg, Bellana, & Bialystok, 2013; cultures, Avis & Harris, 2016; Lee, Olson, & Torrance, 1999; Naito, Komatsu, & Fuke, 1994; Tardif & Wellman, 2000; schizophrenia and autism spectrum disorder, Baron-cohen, Leslie, & Frith, 1985; Frith, 2004; Hughes & Russell, 1993), no research to date has conducted a thorough psychometric

E-mail address: ester.navarro_garcia@tufts.edu.

https://doi.org/10.1016/j.cogpsych.2022.101495
Received 28 January 2022; Received in revised form 27 May 2022; Accepted 7 June 2022
Available online 22 June 2022
0010-0285/© 2022 Elsevier Inc. All rights reserved.

examination of the construct of ToM. Perhaps for this reason, despite the numerous findings regarding ToM, it is still unclear what cognitive processes underly ToM ability. This is partly due to the variability with which researchers operationalize the construct of ToM. For example, ToM has been operationalized as:

- Implicit ToM (i.e., fast, automatic ToM) vs explicit ToM (i.e., slower, deliberative ToM) (Apperly & Butterfill, 2009; van Overwalle & Vandekerckhove, 2009).
- Emergent ToM (based on experience and context) vs latent ToM (expressed as the result of its interaction with general cognitive processes) (Gopnik & Wellman, 1994, 2012; Leslie & Polizzi, 1998; Leslie, 1994).
- Cognitive ToM vs affective ToM (Abu-Akel & Shamay-Tsoory, 2011; Poletti, 2012).
- Empathic ToM vs the ability representing the mental states of others (Preston & De Waal, 2002; Bernhardt & Singer, 2012; van Veluw & Chance, 2014).

The variation in terminology around ToM has contributed to the prolific creation of numerous ToM measurements, however these seem to often present poor psychometric properties. Recent research shows that several measures commonly used to assess ToM likely test different cognitive processes (Warnell & Redcay, 2019; Hayward & Homer, 2017), and it is unclear whether all these processes really tap into an overarching ToM ability (Quesque & Rossetti, 2020). In fact, there has been strong criticism of the way ToM is investigated and conceptually defined for several decades (Bloom & German, 2000; Frith & Happé, 1994), yet the problem persists, as no psychometric studies have been conducted on tasks of ToM. Schaafsma et al. (2015) suggested that one solution to this issue is to avoid treating ToM as a "monolithic" ability. Instead, researchers should consider the possibility of there being different domains of ToM when conducting research studies. However, without proper psychometric work, this goal remains unattainable.

1.1. Theoretical perspectives of ToM

Children's ability to understand mental states (e.g., beliefs, intentions, desires) is a foundational social-cognitive skill related to a variety of healthy developmental milestones, such as social competence, peer acceptance, and academic success (Carlson, Koenig, & Harms, 2013). A vast amount of research has reported that by age 5 there are significant changes in children's understanding of mental states (Harris, 2006; Wellman & Liu, 2004). For example, by the end of their first year, children can treat individuals as agents with intentions (e.g., desires, goals) (Wellman, 2018). Specifically, Brandone and Wellman (2009) found that 6 and 8-month-olds have longer looking times to areas where they expect a person to look for an object than to areas where they do not expect a person to look and Behne, Carpenter, Call, and Tomasello (2005) found that 9–18 month-old infants were more impatient (e.g., reaching, looking away) when an adult could not hand them a toy than when an adult did not want to hand them the toy; this was not true for 6-month-olds. This behavior indicates that infants understand basic intentions by the time they are 9 months, but not earlier.

However, although children can execute many abilities that require basic perspective-taking by the age of 2 (i.e., emotion, intention, or perception), they largely do not understand mental concepts like knowledge and belief, especially between their knowledge and beliefs and the knowledge and beliefs of others (Carlson et al., 2013). This was first demonstrated by Wimmer and Perner (1983), who administered the Sally and Anne false-belief task to children ranging from 3 to 9 years old. While most of the 5–9-year-olds provided accurate responses, the 3–4-year-olds did not, indicating that the ability to represent mental states of other people becomes established at the ages of 4–6. Wellman, Cross, and Watson (2001)'s meta-analysis of 178 false-belief studies reported consistent findings: most 3–4-year-olds do not respond accurately to false-belief tasks compared to older children, indicating that they largely do not understand the mental states of others. Overall, research to date suggests that understanding mental states undergoes a change at age 3–4.

This change is largely distinguished by the difference between Level-1 and Level-2 perspective taking. That is, infants can understand that people see things differently (Level-1 perspective taking), even if they do not yet understand that others can think different things and have different perspectives (Level-2 perspective taking) (Flavell, 1974, 1977; Flavell, Everett, Croft, & Flavell, 1981). For example, Masangkay et al. (1974) administered a series of tasks to 2-to-5-year-olds (e.g., picture task, turtle task) in which objects presented a different perspective for the experimenter and for the children. They found that 2-year-olds correctly indicated when the experimenter could not see an object even when the child could (Level-1), but only older children indicated when the experimenter could see an object from a different perspective (i.e., from the top as opposed to from the left) than the child (Level-1). This suggests that Level-1 develops before Level-2. Similarly, Moll and Tomasello (2006) found that on average 24-month-olds, but not 18-month-olds, helped an adult find an object that was visible to them but not to the adult (Level-1), indicating that children younger than 24 months did not exhibit Level-1 perspective taking.

The developmental differences between Level-1 and Level-2 perspective have been largely taken to support theories within the Competence framework (e.g., theory-theory; Gopnik & Wellman, 1994). That is, children originally have a "theory" of what other people know, but since they are not always correct, they experience communication errors. This forces children to adjust and reconstruct their initial theory to correctly understand what other people know, intend, and believe. Evidence from studies comparing Level-1 and Level-2 perspective taking is thought to indicate that ToM can evolve and become more sophisticated as a result of interaction with the world. However, other findings cannot be explained solely within the Competence framework. Specifically, research has shown that resolution of false-belief tasks is related to executive functioning (EF) performance. This was first reported by Leslie and Polizzi (1998), who examined responses to false belief problems that required more EF, that is, negative false beliefs (i.e., a false belief task where the protagonist's desire is to avoid rather than approach a target). 4-year-olds in the study performed worse in the negative compared to the standard false-belief task, suggesting that more EF was needed for the negative tasks. This was extended

by Carlson and Moses (2001), who examined the relationship between EF (i.e., inhibitory control) and ToM in a sample of preschool-age children. The researchers found that inhibitory control was strongly correlated with ToM performance, even after controlling for factors like language, age, verbal ability, and family size. Numerous developmental and neuroscientific replications of these findings (Gerstadt, Hong, & Diamond, 1994; Van der Meer, Groenewold, Nolen, Pijnenborg, & Aleman, 2011) have led to the conclusion that, unlike the Competence framework suggests, EF is a necessary factor for ToM development, and it likely allows the use of a complex ToM ability. Thus, research examining the relationship between EF and ToM provides support for theories within the Performance framework of ToM (e.g., ToMM theory; Leslie, 1994), that is, children can only utilize their latent ToM correctly when they develop EF naturally with age (i.e., when they are able to inhibit egocentric responses), but not before. The conflict between these theoretical frameworks is known as the competence-performance debate (Wellman et al., 2001; Scholl & Leslie, 2001).

While most of these theoretical frameworks focused on early childhood, research has shown that ToM is critical for many tasks in older children and adults as well, sometimes referred to as advanced ToM. ToM in older age groups has been helpful to further understand aspects of neural correlates of the ToM network (Gallagher & Frith, 2003; Saxe, Carey, & Kanwisher, 2004; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014), individual differences in ToM performance (Dumontheil, Apperly, & Blakemore, 2010; Navarro & Conway, 2021; Navarro, Goring, & Conway, 2021), ToM declines in old age (Rosi, Cavallini, Bottiroli, Bianco, & Lecce, 2016; Wang & Su, 2013) and differences in ToM impairments, (Baron-Cohen, Wheelwright, & Jolliffe, 1997; Baron-Cohen, Wheelwright, Raste, & Plumb, 2001; Chen et al., 2017), among others. For this reason, various current theoretical accounts consider ToM processes in adulthood as well. Most of these theories steer away from dichotic predictions and attempt to describe multiple processes potentially involved in ToM performance. For instance, Apperly and Butterfill (2009) proposed a theory of ToM that accounts for developmental differences throughout the lifespan. Specifically, Apperly and Butterfill's theory suggested that there is a two-system ToM ability that can account for both the EF-ToM relationship and conceptual changes based on experience or context. This dual-system view is based on classical dual-process theory that proposes that human cognition is defined by a distinction between effortless, intuitive, automatic processes (System 1) and effortful, deliberative, operational processes (System 2) (De Neys, 2012; Evans & Stanovich, 2013; Kahneman, 2011; Pennycook, Fugelsang, & Koehler, 2015). According to Apperly and Butterfill, the ToM System 1 is used by infants and young children, but also by adults when the situation does not require effortful processing, such as when there is no perspective conflict. System 1 thus precedes and contributes to System 2, a fully formed ability to comprehend other mental states that requires effortful processing. Tager-Flusberg and Sullivan (2000) also have proposed a two-system view, in which ToM is thought to consist of an implicit social-perceptual component (i.e., implicit attribution of emotions and mental states) and an explicit social-cognitive aspect (i.e., explicit reasoning about beliefs).

Further, building on neuroscientific findings, Schaafsma et al. (2015) proposed that various independent domain-specific low-level processes (e.g., eye gaze, intention tracking) form a ToM construct, instead of having a single-module general ToM. In other words, they claim that ToM is formed by domain-general processes that explain relationships among tasks, but also have domain-specific components that are not accounted for by general processes. Evidence for this account comes from neuroimaging studies showing that ToM is likely not just a single construct (Frith & Frith, 2003; Schurz et al., 2014). Instead, despite general agreement over some of the areas engaged when responding to ToM tasks (i.e., ToM network), recent *meta*-analytic work has found that distinct activation profiles are found when examining separate tasks (as opposed to aggregated tasks) in a brain activation map (Van Overwalle & Baetens, 2009), suggesting that some areas are engaged more often when responding to some ToM tasks, but not others. In addition, brain activation patterns seem to also vary throughout the lifespan, with responses to ToM tasks starting off more diffused in early childhood and becoming more concentrated in adulthood (Bowman, Liu, Meltzoff, & Wellman, 2012; Bowman & Wellman, 2014). This evidence indicates that ToM is likely composed of different processes and can change throughout development.

One way to test some of the predictions made by these theories is to build models of ToM using psychometric approaches. For instance, building psychometric models of a psychological construct can help identify individual difference in task performance that would support the existence of a cognitively effortful System 2 (Apperly & Butterfill, 2009). In addition, non-hierarchical psychometric techniques can help determine whether ToM is indeed better represented by multiple processes or if there is an underlying monolithic ToM (Schaafsma et al., 2015).

1.2. ToM measurement

The first task designed to assess ToM was the false-belief task (Wimmer & Perner, 1983). This was quickly followed by the development of multiple tasks and tests that assess different aspects of ToM. Some of the processes that these tasks measure include false belief understanding (Bernstein, Thornton, Sommerville, 2011; Wimmer & Perner, 1983), accounting for others' perspectives (Dumontheil et al., 2010), the ability to infer mental states from the expression of people's eyes (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Baron-Cohen, Wheelwright, Spong, Scahill, & Lawson, 2001), detection of faux pas (e.g., Baron-Cohen, O'Riordan, Stone, Jones, & Plaisted, 1999), deceptive intentions (e.g., Sebanz & Shiffrar, 2009), understanding others' thoughts (Keysar, 1994), and Level 1 and Level 2 visual perspective-taking (Piaget & Inhelder, 1956; Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010), among numerous others.

Despite the fast proliferation of ToM tasks, psychometric assessments of the validity of ToM measures has been scarce. Psychometric research has been largely employed to validate individual tasks and tests of ToM. For example, Devine and Hughes (2016) tested the validity and reliability of two affective tests of ToM (i.e., the Strange Stories test, Happe (1994); and the Silent Film task (Devine & Hughes, 2013) in a group of children. They found that both a two-factor model and a single factor model with ToM as latent factor fit the data well, suggesting that the items in those tasks have good construct validity. However, the study focused on assessing the tasks items, and did not specify a hierarchical model of ToM or compared models, as their main goal was to validate the tasks, not examine

the validity of a single ToM construct. Overall, the study provides evidence of coherence among tasks that belong to the same ToM domain in children. Gourlay, Collin, Caron, D'Auteuil, and Scherzer (2020) conducted a similar study of the psychometric properties of social cognition in adults, by conducting validity and reliability tests. Exploratory factor analyses suggested that the items in the Strange Stories task and in the Interpersonal Reactivity Index (Davis, 1983) were adequate, indicating that the tasks were reliable, however no measurement models were specified using these tasks to explore an overarching latent variable, possibly because of the limited sample size (i.e., 100). In addition, all models were conducted using varimax rotation and principal component estimation which are not advisable when using data reduction techniques in psychological data, as these techniques tend to assume that data are not correlated, which is not the case in most neuropsychological tests (Kline, 2015). This trend also appears in clinical research; for example, Brewer, Young, and Barnett (2017) reported principal component analyses indicating that test items from one ToM task loaded onto two components in a sample of adults with ASD. However, other research has not been successful at identifying reliable and valid tasks of ToM that load under a same construct. Specifically, Warnell and Redcay (2019) examined the relationship among different ToM measures (including the false belief task, the Reading the Mind in the Eyes test, and pragmatic language comprehension) in children and adults. The researchers found that an exploratory factor analysis did not support a clear structure underlying a ToM factor for the adult group. In addition, even though factor analysis was not possible due to low sample size for the children sample, correlations among the tasks administered to children also revealed poor to no correlations. Similarly, Hayward and Homer (2017) found multiple issues in the reliability and validity of several measures and found no underlying commonality when controlling for comprehension among a group of children. Finally, Chen et al. (2017) found poor reliability and validity across multiple commonly used tasks of advanced ToM (e.g., RMET, Strange Stories) among a group of young adults with schizophrenia. These findings suggest that the measures used to assess ToM might not adequately tap into a single reliable construct.

While most studies have examined the structure of ToM from a methodological perspective, an increasing number of studies are using psychometric techniques to understand underlying processes predicted by theoretical accounts of ToM. Osterhaus, Koerber, and Sodian (2016) specifically attempted to answer the question of whether ToM constitutes multiple processes from a developmental perspective in a group of children. In an EFA, the researcher found that a three-factor solution was the only one that fit the data as opposed to two- and one factor solutions. While the researchers conducted a subsequent CFA to confirm the fit, it is hard to draw conclusions from a CFA that is defined after conducting EFA. Because of its confirmatory nature, the data used to obtain an EFA will fit a CFA almost perfectly (Kline, 2015), therefore conclusions drawn from these CFAs should be taken with caution. In spite of this, the results overall point to the multidimensionality of ToM, as proposed by others (Schaafsma et al., 2015). Furthermore, in a longitudinal study, Białecka-Pikul et al. (2021) also found that ToM domains could be identified as multidimensional processes. Specifically, different perspective taking domains of ToM could be better identified as more than one process (i.e., perspective-taking and perspective tracking), and that the developmental trajectories of those processes were largely heterogenous across ages, contradicting research that only distinguishes between explicit and implicit ToM (e.g., Tager-Flusberg & Sullivan, 2000).

These findings are consistent with recent theoretical accounts proposing that ToM is not likely a single construct, but that instead it might be a composite of both social and cognitive abilities (e.g., Apperly, 2012a, Apperly, 2012b; Gerrans & Stone, 2008). Indeed, neuroscientific studies using fMRI have found clearly distinct activation profiles of the brain areas engaged when using different ToM tasks (Schurz et al., 2014) Thus, it is critical to clarify whether this view of ToM is more consistent with the findings of ToM measures.

One of the first attempts to create a taxonomy (also see Schaafsma et al., 2015) of ToM tasks to raise awareness of the conceptual and terminology issue around ToM dimensions was performed by Quesque and Rossetti (2020). The researchers assessed the face validity of measures of ToM used by researchers from a variety of areas, including developmental, clinical, cognitive, and cognitive neuroscience and concluded that there were apparent differences in the underlying cognitive mechanisms that each of the tasks seemed to measure, including perspective-taking, eye tracking, and inference making. This conclusion indicates that research should focus on identifying and classifying the different subprocesses or dimensions that form ToM, including lower-order cognitive processes, such as kinematic processing (Obhi, 2012), social attention (Heyes, 2014) or emotion recognition (Oakley, Brewer, Bird, & Catmur, 2016) and higher-order processes, such as inhibiting one's perspective, creating models of alternative emotional responses, and updating one's own knowledge. Thus, this review implies that numerous tasks created and implemented to date do not meet the criteria required to measure a single ToM construct.

Despite the conflicting findings regarding construct validity across tasks, many studies have used ToM tasks to examine the relationship between ToM and other cognitive abilities. Specifically, research has shown that ToM is related to verbal ability, executive function, and IQ (e.g., German & Hehman, 2006; Milligan, Astington, & Dack, 2007). In fact, many of the ToM tasks used in the literature have components that, at face value, share processes with constructs commonly studied in the cognitive abilities research literature, such as crystallized and fluid intelligence. To better understand the components of ToM, the relationship among ToM tasks and other cognitive constructs should be examined. This would help elucidate the tasks that measure specific ToM processes, and the tasks that measure other related, but different, cognitive abilities.

More specifically, there is little research examining the relationship between ToM and general intelligence (Spearman, 1904, 1927). General intelligence or *g* is thought to be the result of the positive manifold, that is, the largely replicated finding that cognitive abilities are consistently positively correlated. General intelligence research is related to various specific cognitive abilities. One of the earliest models of intelligence was the fluid/crystallized (Gf/Gc) theory (Cattell, 1963, 1971; Horn, 1994). The Gf/Gc model proposed that general intelligence was the result of two specific and exclusive abilities: Gf or fluid intelligence and Gc or crystallized intelligence. Gf is the ability to solve problems in novel situations, regardless of previous knowledge and Gc is the ability to solve problems using previously acquired skills, largely related to the amount of formal schooling one has been exposed to (Kan, Kievit, Dolan, & van der Maas, 2011). These two abilities have been expanded and incorporated into more recent models of intelligence, including the Cattell–Horn–Carroll (CHC) model (McGrew, 2009), which combines the Gf/Gc model with other specific abilities, such as visual-spatial

(Gv), processing speed (Gr) and memory retrieval (Gr). Importantly, Gf and Gc remain two of the strongest factors in models of intelligence and have been replicated consistently across the literature and across neuroscientific and developmental studies.

Research has found that ToM is related broadly to IQ (Baker, Peterson, Pulos, & Kirkland, 2014; Dodell-Feder, Lincoln, Coulson, & Hooker, 2013). Some researchers have proposed that this relationship may be the result of overlapping processes between ToM and intelligence (Coyle et al., 2018), suggesting that ToM tasks are largely just measuring intelligence. While this hypothesis has been debated (Navarro et al., 2021), the lack of psychometric research on ToM has hindered further progress on this issue. Understanding the similarities and differences between ToM and intelligence is necessary to fully comprehend the extent to which ToM tasks assess the same underlying ability above and beyond general cognitive ability. In addition, theoretical accounts are increasingly suggesting that ToM tasks are most likely tapping into multiple components of a ToM ability (Schaafsma et al., 2015). Psychometrically assessing ToM in comparison to tasks of general intelligence could shed light in both of these open questions.

1.3. The current study

The goal of this study was to provide the first psychometric examination of ToM tasks by comparing performance on the tasks to measures of fluid intelligence (Gf) and crystallized intelligence (Gc) (Cattell, 1963) and to explore the claims made by theoretical accounts that propose that ToM should not be considered a monolithic construct (Apperly & Butterfill, 2009; Devaine et al., 2014; Schaafsma et al., 2015). Examining the differences between these constructs would allow us to explore whether the processes tapped by ToM tasks represent a unique ToM ability, or whether they instead measure multiple components. For this purpose, participants completed a battery of ToM, Gc, and Gf tasks. Confirmatory factor analyses were used to examine the structure of the data. We hypothesized that if ToM tasks represent a monolithic cognitive ability, then a three-factor model should best fit the data.

In addition, a psychometric network modeling analysis was conducted to examine the relationship among ToM, Gc, and Gf tasks. Psychometric network modeling conceptualizes cognitive abilities as interconnected networks composed of interactive processes (see Epskamp & Fried, 2018). Psychometric networks are a powerful visualization tool to explore anticipated or unknown relationships amongst variables in a dataset and, unlike latent variable modeling, they are not constrained by the principle of local independence (i.e., the assumption that a latent factor causes any and all covariation among measures of the same construct). In addition, network modeling can account for the one-to-one relationships amongst tasks belonging to the same construct while at the same time estimating individual relationships between tasks belonging to different constructs.

2. Method

2.1. Design and participants

An online sample of 208 participants was recruited using Amazon's Mechanical Turk (MTurk). The number of participants is based on the minimum sample size required for a three-factor Confirmatory Factor Analysis (CFA; Wolf, Harrington, Clark, & Miller, 2013). All participants were located in the US and were over 18 years of age. Participants' ages ranged from 18 to 69 years old ($M = 39.89$; $SD = 9.34$, Median = 39). 116 participants identified as female. In terms of ethnicity, 148 participants identified as Caucasian, 13 identified as Black/African American, 9 identified as Asian, 3 identified as Hispanic/Latino, and 7 identified as mixed ethnicity. 23 participants did not report ethnicity. None reported being color blind. In addition, all participants reported having correct-to-normal vision and were fluent in English. The final sample size after multivariate outliers were removed was $N = 203$.

The design of the study was a correlational approach using two different psychometric modeling techniques. To conduct factor analyses, it is recommended that each latent construct includes at least three tasks. In this study, participants completed 9 tasks in total: 3 tasks of ToM, 3 tasks of Gf, and 3 tasks of Gc. Participants were randomly assigned to complete the tasks in one of three different orders. In order 1 ($n = 74$), participants first completed the Gf tasks, followed by the Gc and by the ToM tasks. In order 2 ($n = 58$), participants first completed the ToM tasks followed by the Gf and Gc tasks. In order 3 ($n = 76$), participants first completed the Gc tasks followed by the ToM and Gf tasks[1].

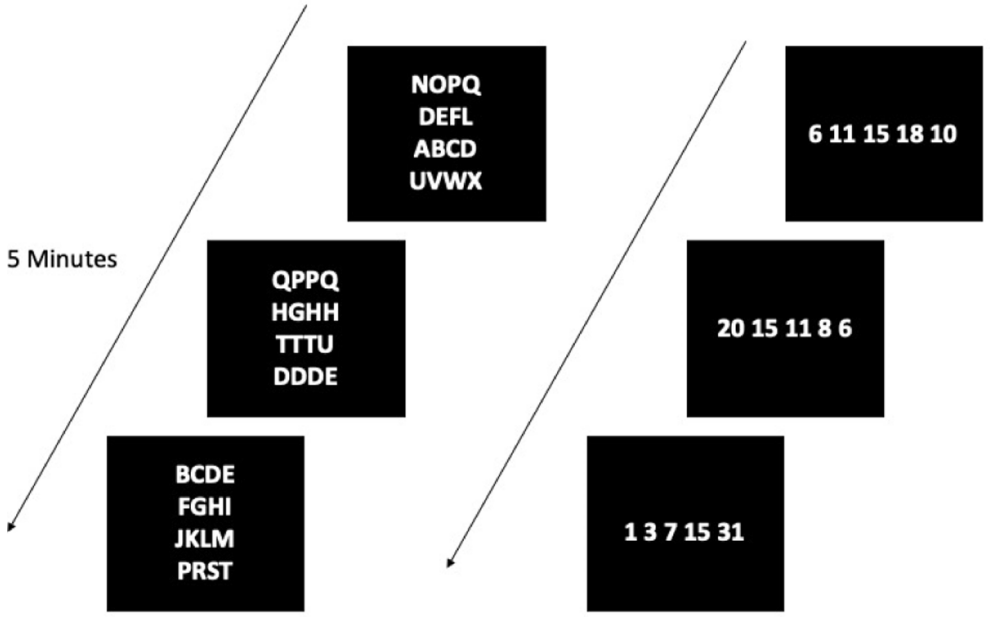
3. Measures

3.1. Theory of mind tasks

As discussed above, there are multiple tasks of ToM available. In this study, we decided to select three tasks based on different criteria. First, we selected measures that have been previously validated and tested un adult samples to avoid ceiling effects. Second, we selected measures from different domains of ToM with the goal of achieving as much variability as possible within the ToM construct based on the taxonomy described by Quesque and Rossetti (2020). Finally, we selected measures that were commonly used in the literature while also considering the criteria required for ToM tasks: nonmergent (i.e., a task requires representing the mental state of another person that differs from the participant's mental state) and mentalizing (i.e., success in a given task necessitates understanding others' mental states and not be attributed to lower-order cognitive processes) criteria (Quesque & Rossetti, 2020). The ToM

[1] No significant differences were found in any of the dependent variables based on order of administration.

measures that were used in the study involve (a) taking the perspective of another person (i.e., director task), (b) inferring mental states from people's eyes (i.e., Reading the Mind in the Eyes), and (c) interpreting mental states in socially inappropriate situations (Short Stories Questionnaire). Each task is thought to represent different dimension of ToM, visual perspective-taking (director task), social cognitive (Short Stories Questionnaire) and social perceptual (Reading the Mind in the Eyes). The tasks lasted approximately 15 min.



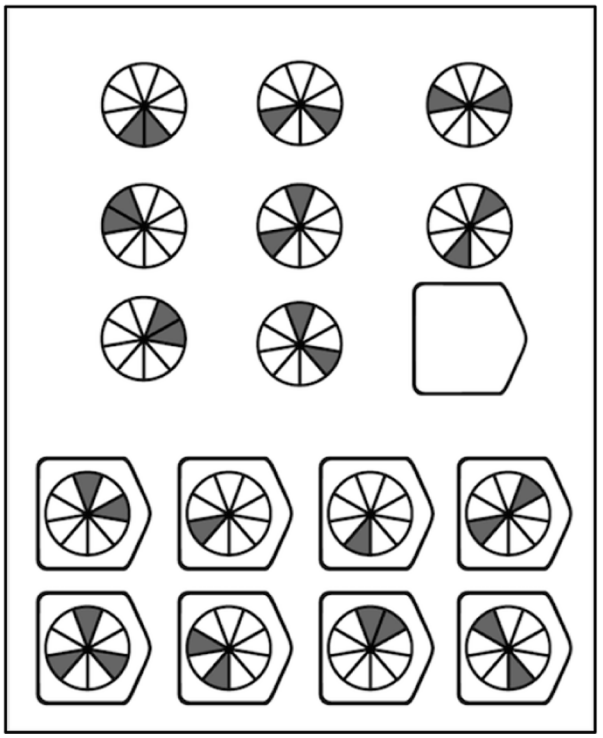


Fig. 1. Sample trials of tasks of fluid intelligence. Letter series (top left), number series (top right), Raven's progressive matrices (bottom).

Director task. The task was proposed by Keysar, Lin, and Barr (2003) and automated by Dumontheil et al. (2010). The current version was an automated version adapted from Legg, Olivier, Samuel, Lurz, and Clayton (2017). The task includes two conditions (Director, No Director) and 2 trial types (Experimental, Control). The stimuli are set up in a 4 × 4 shelf containing eight different objects arranged in different positions. In the Director Condition, an avatar (the Director) stands behind the shelf. Some of the compartments in the shelf are occluded from the Director's view so that only the participant can see those objects. The participant is asked to attend to the instructions that the Director gives them via a speech box. On each trial, the Director asks the participant to select one of the objects in the shelf (e.g., "the small cup"). The participant responds by clicking on the correct object within the shelf. Participants were asked to consider the perspective of the Director when responding to instructions. This is thought to assess theory of mind because the participant has to remember that the perspective of the Director is not the same as theirs. In the *No Director* condition, participants are shown the same shelf, but the Director is not behind it anymore. Instead, participants are told to ignore all objects placed in the slots with red backgrounds. This condition does not require theory of mind and instead requires the participant to inhibit prepotent information while keeping in mind a rule, therefore just requiring general executive function. Both conditions present experimental trials (trials containing a target and a competing distractor that can be the most appropriate response but only from the perspective of the participant), control trials (trials containing no competing distractor) and filler trials. Control and experimental trials appear in a pseudorandom intermixed order throughout the task and the order of presentation of the stimuli is counterbalanced across participants. Participants respond to a total of 16 control trials, 16 experimental trials and 64 filler trials in each condition. Participants also complete a practice trial before the Director condition. Participants have 5 s to respond to each instruction. Correct responses to the experimental trials of the director condition were calculated.

Reading the Eyes in the Mind (RMET). The RMET task (Baron-Cohen et al., 2001) was programmed on a computer. In the task, participants are presented with a series of 36 black and white photographs of the eye region of the face of females and males of different ages. One photograph is presented at a time and participants have no time limit to respond. Four words describing the potential emotion conveyed by the eyes are presented together with the photograph. Participants must select the word that best describes what the person in the photograph is feeling (e.g., sad, happy, scared, depressed). A glossary of the tasks words was accessible to the participants via link. The test is thought to assess how well a person can understand other people's mental states. Responses were self-paced. RMET scores range from 0 to 36 in a discrete fashion. Accuracy is recorded. The task lasts approximately 10 min.

Short Stories Questionnaire (SSQ). SSQ (Dodell-Feder et al., 2013; Lawson, Baron-Cohen, & Wheelwright, 2004) was implemented on Qualtrics. The test contains 10 short stories, each divided into three sections. The stories involve utterances made by a character that could upset another character in the story. In this task, participants must infer the mental states of the characters (i.e., how they felt, what they thought). There are a total of 30 sections with at least four utterances in each section. 10 sections contained *blatant* target utterances (e.g., incorrectly estimating a middle-aged woman's age), 10 contained *subtle* target utterances (e.g., lying about remembering someone's name) and 10 contained *filler* control utterances (e.g., discussing the weather). Each section contained a question asking the participant whether something said in the story could have upset someone and to indicate what part of the text corresponded to the upsetting utterance. Each of the 10 stories included a filler question. The order of presentation of the stories was randomized. Responses were self-paced. The number of correctly identified targets was calculated. Scores range from 0 to 20 in a discrete fashion. The task lasts approximately 15 min.

3.2. Fluid intelligence tasks

The tasks are thought to measure the ability to follow rules and solve novel problems.

Letter Series. In the letter series task (Ekstrom, Dermen, Harman, 1976), ten trials of four letters are presented on the screen, with one trial at a time. All trials present four series of letters that followed a certain pattern except for one set. To respond correctly, participants must select the letter set that does not follow the pattern. After responding, the next set of letters appears on the screen. There is no time limit per trial (see Fig. 1). Accuracy is measured in this task. Participants were given three practice trials before completing the task. The task automatically ends after 5 min.

Number Series. In the number series task (Thurstone, 1938), ten trials shown one at a time are presented showing a series of numbers of varying lengths in it. Each series of numbers is organized following a specific rule or pattern. Participants are asked to select the number that would be consistent with the series from five given choices for each trial (see Fig. 1). Participants were given three practice trials before completing the task. Accuracy is measured in this task. The task automatically ends after 5 min.

Raven's Progressive Matrices. The advanced version of Raven's figural inductive reasoning task (Raven, 1938) was used in this study. Participants completed a total of 18 items (Hamel & Schmittmann, 2006). Participants were randomly assigned to one of two task orders (odd trials or even trials). Each item is part of a pattern of eight black and white figures arranged in a 3 × 3 matrix in which the last bottom-right figure is missing (see Fig. 1). At the bottom of the matrix is a list of eight possible figures to choose from. Only one of those figures is the correct answer that best completes the pattern of the missing piece in the matrix. Each item follows a series of rules that the participant needs to find and keep in mind to find the right answer. Participants were given three practice trials before completing the task. Total correct responses were calculated. The task ends automatically after 15 min.

3.3. Crystallized intelligence tasks

All tasks are thought to measure previously acquired knowledge.

Synonyms. The synonyms test (Woodcock-Johnson Battery of Cognitive Ability III, 2001) presented participants with 10 words shown one at a time each with a list of possible answer choices. Participants had to choose the word whose meaning was the same as

the initial word displayed on the screen. Accuracy is measured in this task. Participants had 5 min to answer all 10 questions.

Antonyms. The antonyms test (Woodcock-Johnson Battery of Cognitive Ability III, 2001) is identical to the Synonyms test, with the exception that participants have to choose from a list of options the word that represents the opposite meaning to the word displayed. Accuracy is measured in this task. Participants have 5 min to answer all 10 questions.

General Knowledge. The test (Woodcock-Johnson Battery of Cognitive Ability III, 2001) consisted of 10 questions regarding general knowledge (e.g., "What planet is furthest from the sun?"). Participants have to type out their answers to respond and are asked to enter "I don't know" if they do not know the answer. Accuracy is measured in this task. Participants have 5 min to answer all questions.

3.4. Procedure

All tasks were administered via Qualtrics and participants accessed the study from Amazon's MTurk. Participants were assigned to one of three counterbalanced orders as indicated above. Tasks within each construct were presented always in the same order. Participants were allowed breaks in between tasks. Completing the battery of tasks took approximately 90 min. Participants were compensated with $15.

4. Results

As a reminder, responses to experimental trials of the director task and SSQ were included in the analyses[2]. Descriptive statistics for each measure and reliability estimates are presented in Table 1. All but one measure demonstrated adequate reliability as measured by Cronbach's alpha (i.e., $\alpha \geq 0.70$; Nunnally, 1970) for each of the given constructs. All ToM tasks presented the lowest reliability, just around the minimum value of 0.70. Bivariate correlations between measures are reported in Table 2. Gf and Gc tasks were moderately or strongly correlated. Gc and Gf measures were also correlated with each other as it was expected based on models of intelligence. The ToM measures presented less consistent correlations, as previous research has shown (e.g., Coyle et al., 2018; Warnell & Redcay, 2019; Hayward & Homer, 2017). In addition, all tasks appeared to correlate with measures of Gf. For example, the director task and SSQ presented low but significant correlations with all tasks, not just with ToM tasks, and the RMET seemed strongly related to the Gf measures in particular. We further explored these relationships using confirmatory factor analyses. Data cleaning procedures can be found at: https://osf.io/tpnj9/.

4.1. Confirmatory factor analysis

Confirmatory factor analysis is a technique used to test and estimate relationships among observed and unobserved variables to construct a measurement model. Fit indices based on Kline (2015) were followed: chi-square to degrees of freedom ratio lower than 2, a Comparative Fit Index (CFI) greater or equal to 0.90, a Standardized Root Mean Square Residual (SRMR) lower or equal to 0.08, and a Root mean square error of approximation (RMSEA) between 0.05 and 0.10 (Kline, 2015). In this study, CFA was used to assess the construct validity of the tasks by comparing model fit and path loadings. CFA requires the use of an estimation algorithm to compare iterated sets of values with the goal of minimizing the difference between the observed and the implied correlation matrix. Robust maximum likelihood (Gibson & Ninness, 2005) is an adequate estimator for data that present multivariate non-normality (as was the case in these data) and was used in this study. Data from N = 203 participants were used.

Three models were specified. The first model, Model 1, was a one-factor model where all manifest variables were predicted by a single general construct. Model fit indices are in Table 3.

Model 1 presented poor fit based on Kline's fit indices, with no indices within standard ranges. While the fit presented a poor model, the standardized factor loadings were overall adequate, and only Raven's Progressive Matrices presented loadings under 0.30 (see Fig. 2). This indicates that, as expected, a model with a single factor does not adequately represent the ability that the measures are thought to assess.

Model 2 was conducted next to examine whether the ToM tasks would be better represented by a Gf factor compared to a separate factor, as some researchers have proposed (Coyle et al., 2018). Model 2 was a two-factor model where Gf and Gc were the latent factors. The tasks corresponding to the traditional ToM and Gf constructs were combined in this model based on their bivariate correlations. Fit indices for Model 2 are in Table 3. Overall, Model 2 did not present an excellent fit based on Kline's fit indices, however most of the indices were close to good fit. Compared to Model 1, Model 2 presented better fit indices across the board. Standardized factor loadings in Model 2 were adequate for the Gc and Gf factors, despite the fact that the ToM measures were loaded into the Gf factor. The correlation between Gf and Gc was strong (see Fig. 2). These findings seem to indicate that although a two-factor solution was not a perfect fit for the data, nevertheless Model 2 seemed overall better than Model 1 and was not a complete misrepresentation of the data. Finally, Model 3 was conducted to examine whether a theoretically-driven three-factor model provided a more adequate representation of the data.

Model 3 was a three-factor model where each set of tasks was grouped under the psychological construct they represent

[2] Group effects of the director task were examined by comparing control and experimental trials of the director task condition. As expected, participants performed significantly worse in experimental trials compared to control trials (t (2 0 2)= −13.24, p =<.001 (experimental $M = 0.59$, control $M = 0.94$), Cohen's $d = -1.26$.

Table 1
Descriptive Statistics.

Variables		Latent Construct	*N*	*M*	*SD*	Skew	Kurtosis	Range	*α*
1. Ravens Progressive Matrices	RV	Fluid Reasoning (Gf)	203	9.28	384	−0.89	−0.78	0–18	0.81
2. Letter Series	LS		203	6.77	2.59	0.05	−0.40	0–10	0.77
3. Number Series	NS		203	9.40	3.07	−0.26	−0.61	0–15	0.85
4. General Knowledge	GK	Crystallized intelligence (Gc)	203	7.09	2.26	−9.9	0.34	0–10	0.79
5. Synonyms Task	SYN		203	5.99	2.52	−0.63	−0.33	0–10	0.86
6. Antonyms Task	ANT		203	6.16	2.15	−0.48	−0.68	0–10	0.85
7. Director Task	DT	Theory of Mind (ToM)	203	0.59	0.39	−0.48	−1.61	0–1	0.66
8. Reading the Eyes in the Mind	RME		203	30.14	4.02	−0.92	1.12	0–36	0.72
9. Short Stories Questionnaire	SSQ		203	10.15	3.33	−0.20	−0.24	0–20	0.72

Table 2
Correlations among variables.

Variable	1	2	3	4	5	6	7	8
1. General Knowledge	–							
2. Synonyms	0.50	–						
3. Antonyms	0.43	0.66	–					
4. Letter Series	0.17	0.19	0.18	–				
5. Ravens	0.52	0.38	0.44	0.36	–			
6. Number series	0.39	0.33	0.34	0.49	0.56	–		
7. SSQ	0.18	0.15	0.15	0.16	0.33	0.14	–	
8. RMET	0.33	0.32	0.32	0.30	0.41	0.39	0.24	–
9. Director Task	0.16	0.16	0.16	0.18	0.32	0.27	0.18	0.17

Note. All correlations were significant at $p = <0.05$.

Table 3
Model Fit Indices for All CFA Models and NMA.

Fit Indices	χ^2	*df*	x^2/df	CFI (TLI)	RMSEA	SRMR
Recommended fit (Kline, 2015)			**≦2**	**≧0.90**	**≦0.08**	**0.05–0.10**
Model 1: One predictor	164.19	28	5.86	0.73 (0.65)	0.16	0.149
Model 2: GF + TOM	84.92	27	3.15	0.88 (0.84)	0.103	0.109
Modified Model 3: GF + GC + TOM	79.62	25	3.18	0.89 (0.84)	0.104	0.106
Network model	21.46	11	1.95	0.98(0.92)	0.068	0.033

Note. It is common that the Network model presents excellent fit, nevertheless, because the network model is an exploratory analysis, it is not possible to directly compare it to the CFAs.

theoretically. Model 3 presented two covariances for the latent factors with values over 1, indicating that the model was misspecified. This was possibly due to the weak correlation among ToM tasks. In addition, contrary to what was expected, fit indices indicated that the model did not present an excellent fit to the data. Unlike the Gc and Gf tasks, the ToM tasks all presented low loading paths and the correlations between the ToM factor and both the Gc and Gf factors showed correlations above 1, suggesting that perhaps the tasks in the ToM factor might have overlapping variance with some of the tasks in the other factors.

To further understand this, modification indices were used to re-examine Model 3. Modification indices are estimates of the amount by which the chi-square value of a given model would be reduced, and therefore fit increased, if a specific parameter were modified in the model. That is, modification indices allow researchers to understand the ways in which the model fit could improve based on a data-driven approach. Because of this, it is not advisable to use modification indices to specify a model a priori, but rather to examine potential issues in the existing model a posteriori. Modification indices for Model 3 suggested that the fit of the model would improve if the RMET were predicted by both Gf and ToM (see Fig. 2). This modification considerably improved the fit of the model (see Table 3) and the manifest variables loadings of the ToM latent factor, as well as avoiding covariances over 1. This suggests that the RMET is contributing variance to both constructs and therefore that the task assesses shared processes across these two constructs. The fit indices of Model 3 were close to those of Model 2, with only a slight improvement.

In general, none of the models presented excellent fit according to the fit indices. Overall, the modified Model 3 presented the most adequate indices, however there was no difference with the other models.

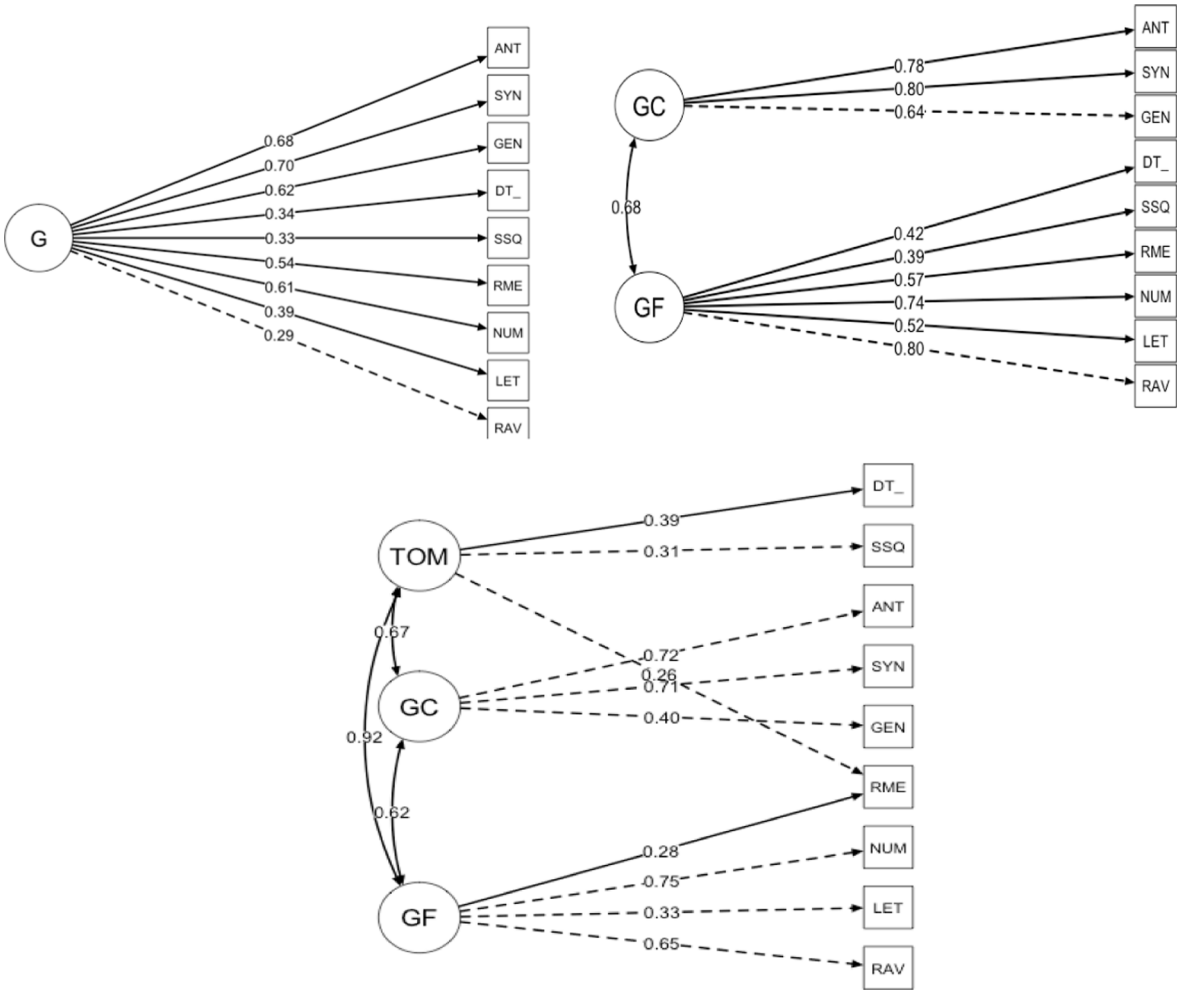


Fig. 2. Model 1 (top left): One-factor model. Model 2: Two-factor model (top right). Model 3: Three-factor model with modifications (bottom). Standardized factor loadings are presented. All loadings were within adequate range (>0.30) with the exception of Raven's. ANT = Antonyms, SYN = Synonyms, GEN = General Knowledge, DT = director task, SSQ = Short stories, RME = Reading the eyes in the mind test, NUM = Number series, LET = Letter series, RAV = Raven's progressive matrices. None of the models presented excellent fit. Model 1 presented the worst fit. Model 3 presented adequate fit but was not significantly different from Model 2.

4.2. Exploratory factor analysis

Given the lack of fit of the CFAs, we decided to conduct an exploratory factor analysis (EFA) to understand whether the data were indeed a good representation of the measurement model constructed in the CFAs. A parallel analysis indicated that 2 factors should be retained, rather than 3. This suggests that the third factor was likely so small that it was little more than random noise. The rotation estimator was Oblimin[3], given the correlations among the variables. All variables with loadings greater than 0.30 were considered to load on a given factor. When 2 factors were retained, the ToM tasks were loaded under the same factor as the Gf tasks, mirroring the results of the Model 2 CFA. To further explore the data, we also conducted an EFA retaining 3 factors following the theoretical framework. The results showed that all measures of Gc loaded adequately under the same factor. However, the director task and RMET loaded under the Gf factor with the rest of the Gf measures, whereas the SSQ was almost entirely loading on the third factor by itself. These results follow those found by Warnell and Redcay (2019) and suggest that the tasks have common processes with fluid reasoning. We next conducted an exploratory network model to better understand the relationships among individual measures.

[3] Extraction techniques produce factors that are orthogonal and atheoretical. Rotation allows the transformation of the factor loadings, so they become more interpretable. Oblimin is an oblique (as opposed to orthogonal) extraction technique, therefore it allows the factors to be correlated (which is often the case in psychological studies).

4.3. Network model analysis

Exploratory Network Model Analysis (NMA) is an alternative analysis that conceptualizes cognitive abilities as interconnected networks composed of interactive processes (see Epskamp & Fried, 2018). In this technique, observed manifest variables are represented by *nodes* and estimated partial correlations amongst them are modeled via connections called *edges*. Therefore, this technique does not need the assumption of a superordinate unobservable factor. NMA can be used in conjunction or as an alternative to latent variable modeling and it presents a number of benefits. For example, because of its exploratory nature, NMA can be used as a powerful visualization tool to explore anticipated or unknown relationships amongst variables in a dataset. NMA is also not constrained by the principle of local independence, unlike traditional latent modeling. The principle of local independence assumes that a latent factor causes any and all covariation among measures of the same construct, and therefore CFA does not allow to observe the variance that manifest variables potentially have in common. Instead, NMA estimates associations between observed variables without assuming that a latent cause is responsible for any and all covariation among measures of the same construct. For this reason, NMA can account for one-to-one relationships among nodes belonging to the same construct while at the same time estimating individual relationships between nodes belonging to different constructs. Finally, NMA estimates associations between all observed variables, therefore it is ideal for modeling cognitive theories that propose overlapping processes among the same construct. In addition, since NMA is an exploratory technique, it can be used on the same data as the CFA.

In this study, NMA was used to examine tasks that assess ToM, Gf, and Gc and the extent to which they are related among them. NMA was conducted on the correlation matrix extracted from the same 203 participants. The model was conducted and visualized using the *qgraph* package in R. The method and techniques used in this study are consistent with recommendations from the network modeling tutorial written by Epskamp and Fried (2018). To conduct an NMA, two parameters must be set. Gamma is a hyperparameter that determines whether the model favors a more simple or complex structure per the number of estimated edges. Lambda is a tuning parameter that determines the rigorousness of removal of identified spurious edges that occur due to sample error. The NMA was generated using the graphical least absolute shrinkage and selector operator (gLASSO) regularization method to determine the level of network sparsity. Specifically, the extended BIC method was utilized, which produces simpler models, as gamma is automatically set to its most conservative setting (0.50). Consistent with Epskamp, Lunansky, Tio, and Borsboom (2018), lambda was set to remove spurious (false-positive) edges while at the same time maintaining as many true edges as possible (i.e., 0.01). The settings used for the network model are designed to facilitate high-specificity during the estimation process, and high-sensitivity regarding network edge-pruning.

Fig. 3 shows the results of the network model. First, both Gf and Gc measures show strong partial correlations and form two closely related but independent constructs. One of the tasks, Raven's, seems to have an especially central position in regards to the correlations among all three psychological constructs. While the Gf and Gc tasks cluster together, the ToM tasks do not seem to represent a strong

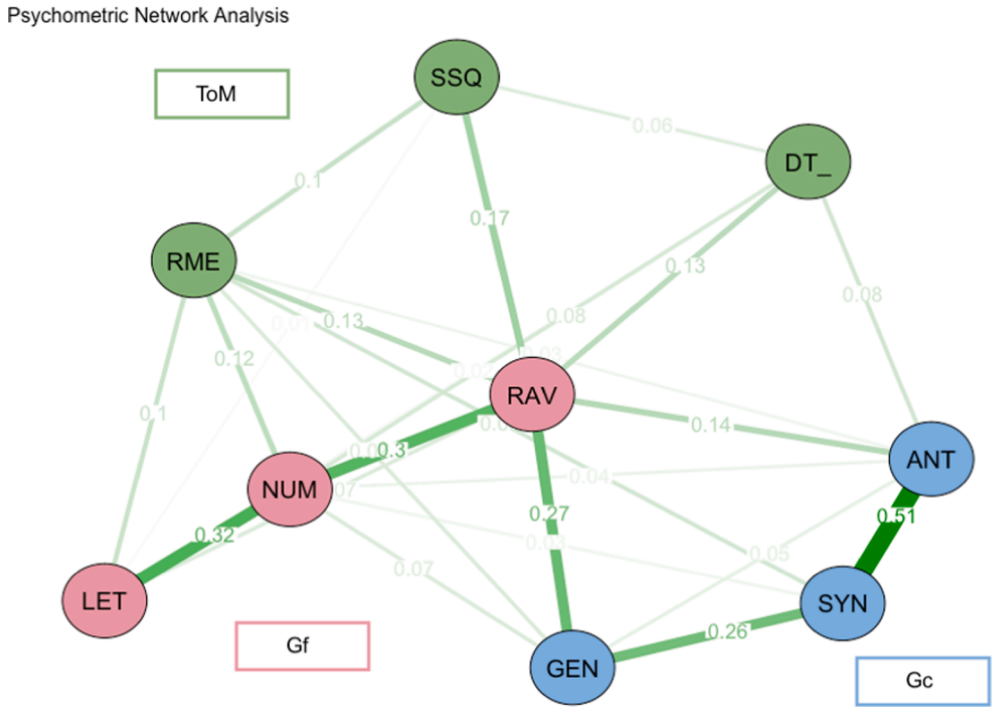


Fig. 3. Network model. Nodes represent the tasks measured in the study. Edges represent the partial correlations among measures. Colors represent the theoretical construct they assess.

unified cluster. Similar to the findings observed in the CFA, in the NMA the ToM tasks are more sparsely related to each other than the tasks that form the other constructs. In fact, they seem more related to other non-ToM tasks. Specifically, SSQ is slightly more related (0.17) to Raven's Progressive Matrices than it is to either of the other ToM tasks (0.1 and 0.06, respectively). In addition, the director task and the RMET do not share any significant edges with each other, despite loading on the same latent factor in the EFA, suggesting that the relationship between the two measures that were observed in the EFA might be due to their relationship to Raven's. Moreover, RMET seems to be more closely related to all the Gf tasks than to any other task, clustering closer to the Gf construct, rather than the ToM construct, following the findings of the EFA. Overall, the ToM tasks do not seem to form a uniform construct, and rather seem to share processes with the tasks belonging to the other two constructs.

The results of the NMA show that the three ToM tasks measured in this study are not as strongly related to each other as previously thought, thus questioning the monolithic view of ToM as an overarching construct. In addition, these findings replicate recent research suggesting that there is little coherence among ToM tasks (Schurz et al., 2014: Warnell & Redcay, 2019). Although the ToM tasks used in this study might be tapping on to *some* dimension of a ToM construct, the findings above suggest that there are clear differences in the processes the tasks assess and they possibly also measure other cognitive abilities (such as Gf), in addition to solely ToM. This indicates that more psychometric research is necessary to understand what tasks should be used to assess ToM in adults, but also to understand what components of ToM the tasks measure.

5. Discussion

This study constitutes the first psychometric assessment of the structure of ToM. The goal was to compare ToM to crystallized intelligence (Gc) and fluid intelligence (Gf) to (a) understand whether ToM tasks measure the same construct or whether they assess different components of ToM, and (b) explore whether the tasks assess ToM ability, above and beyond general cognitive ability. More broadly, the study contributes to novel theoretical approaches that view ToM as a multidimensional construct, rather an as a monolithic ability (e.g., Apperly & Butterfill, 2009; Devaine et al., 2014; Schaafsma et al., 2015).

As previous research has indicated, the ToM measures were poorly correlated (Warnell & Redcay, 2019), and presented average to low reliability, indicating that internal consistency can be improved. The CFA showed that none of the measurement models presented excellent fit. Specifically, Model 2 (the two-factor model with a Gf-ToM latent factor and a Gc latent factor) presented similar indices and path loadings to Model 3 (the model with Gf, Gc, and ToM). In addition, Model 3 had to be modified a posteriori based on modification indices due to the model being misspecified. These findings suggest that the tasks used to measure ToM have common processes with tasks of Gf as others have found (Coyle et al., 2018, Navarro et al., 2021). In fact, the modified CFA model showed that the RMET shares processes with Gf and that a model where RMET was loaded under both ToM and Gf improved model fit. This was further supported by the exploratory factor analysis (EFA) in which the RMET and the director task loaded under the Gf factor, whereas the SSQ loaded separately, indicating that the ToM tasks tested here do not represent a unified construct. Finally, the NMA presented a description of the individual tasks. Specifically, the network model showed that measures of the well-established Gc and Gf constructs presented strong edges among the corresponding tasks and overall clustered together. However, ToM tasks were sparsely related, with weak edges. The nodes were also closer to the Gf tasks (especially Raven's), than to each other. Specifically, the RMET seemed related to all Gf tasks but only presented a weak edge with the SSQ, and no edges with the director task. Similarly, the director task shared edges with measures of Gc and with the SSQ but not with the RMET, whereas SSQ presented weak edges with both ToM tasks and with Raven's. Overall, these findings suggest that at least in samples of neurotypical adults, some of the most common measures of ToM do not adequately represent a single construct.

More broadly, the current findings provide support for existing theoretical frameworks of ToM. Specifically, the finding that multiple tasks representing different domains of ToM form a latent construct (even if it can be improved), supports the claims made by Schaafsma et al. (2015). That is, instead of a single ToM, there can be multiple domain-specific processes that are related to each other but represent different aspects of social cognition. While this study did not include an exhaustive battery of ToM domains, the findings in this study show that different identified domains of ToM are related to each other and are distinguishable from, albeit related to, general cognitive ability. In addition, the domains represented by these tasks present psychometrically divergent properties that make it difficult to describe a single latent factor. Instead, based on Schaafsma and colleagues' model, the NMA could be a better depiction of domains of ToM. In subsequent studies, more tasks of each domain should be examined to confirm these preliminary results. Importantly, the theory proposed by Schaafsma and colleagues' supports the use of psychometric network models as an alternative to hierarchical latent variables. As discussed, one of the main goals of this study was not only to offer a latent variable analysis of ToM, but also to provide evidence for the benefits of network modeling as a non-hierarchical conceptual framework of ToM. That is, creating a model where an overarching latent construct is not required. As mentioned, this is a prediction of Schaafsma et al. (2015), and it is also in line with current trends in psychometrics research arguing for the need to represent psychological data in non-hierarchical structures that better represent neural networks (e.g., Kovacs & Conway, 2016; Duncan et al., 2010).

Furthermore, the current findings provide some support for the two-system theory described by Apperly and Butterfill (2009). Specifically, that individual differences in ToM when conflict arises indicate individual variability. Individual variability in ToM tasks performance is a requirement for system 2 of the dual-process theory. Specifically, an effortful mechanism requires the engagement of cognitively effortful processes to minimize errors in performance while dealing with high cognitive load introduced through conflict. Thus, variance among the tasks indicates variability in performance when responding to conflicting ToM tasks. This also aligns with some of the tenets proposed by the ToM-Mechanism theory (Leslie, 1994; Leslie, German, & Polizzi, 2005; Leslie & Polizzi, 1998). For instance, in this study, ToM was especially related to Gf, which is strongly correlated to executive functions, like working memory. The finding that ToM and general cognitive ability are related has been previously found in the literature, but has not been observed while

modeling constructs with multiple tasks. Just like cognitive ability, ToM is developed in early childhood and increases exponentially throughout the life span, gradually developing individual differences. The current findings provide support for the intrinsic relationship among these abilities in adulthood.

This study also does not find support for recent research that proposes that ToM is just a byproduct of general intelligence (Coyle et al., 2018) by showing that although there is a clear contribution of general intelligence to ToM, this is a separate construct from Gf and Gc.

While describing the exact mechanisms underlying these findings is beyond the scope of the paper, it is possible that the relationships among ToM and Gf tasks is due to a divergence among ToM tasks and a convergence among ToM-Gf tasks. In other words, each of the ToM tasks tap different ToM dimensions (i.e., social affective, social perceptual, perspective-taking), whereas they all require taxing fluid reasoning to some extent. Therefore, it is possible that these subdimensions of ToM are unrelated enough that it prompts weaker correlations, but they are still dependent on general cognitive ability. This indicates that it would be beneficial for the field to conduct further research on the components of ToM rather than a single ToM ability. That is, just like much research is dedicated to identifying and understanding components of general intelligence (e.g., processing speed, long term memory retrieval, visuospatial ability), ToM research would benefit from a similar identification and study of its components. Indeed, as mentioned in the review, there are discrepancies in the neural underpinnings of ToM as shown by studies using fMRI. In a meta-analysis, Schurz and colleagues (2014) found that when activation loci were broken down by task, there were clear differences in the activation profiles of the groups. This indicates that different tasks are indeed activating different regions and that using tasks indistinctly to measure ToM can obscure potential differences in the neural and cognitive mechanisms engaged across ToM dimensions. In addition, the relationship between ToM and Gf could be partially the result of the strong relationship between ToM and executive function (German & Hehman, 2006; Milligan et al., 2007) and Gf and executive function, especially working memory (Kane & Engle, 2002; de Abreu, Conway, & Gathercole, 2010; Unsworth, Fukuda, Awh, & Vogel, 2014). Specifically, the relationship between ToM and Gf has been thoroughly documented in the literature, prompting some researchers to suggest that ToM and executive function are intrinsically related (Carlson, Moses, & Breton, 2002; Carlson, Moses, & Claxton, 2004; German & Hehman, 2006; Perner, Lang, & Kloo, 2002; Hala, Hug, & Henderson, 2003) since they both share underlying properties and developmental changes and both tend to appear at around the same age (Carlson, 2005; Garon, Bryson, & Smith, 2008). On the other hand, the well-documented relationship between executive function and fluid intelligence is thought to be due to core processes, such as individuals' ability to hold information in mind while avoiding interference. These similarities could be behind the strong relationship between ToM and Gf in this study.

Moreover, this research is in line with previous studies showing that ToM tasks are related to intelligence, however the construct of ToM seems distinct enough from intelligence to be considered a separate ability (Navarro et al., 2021). While research on ToM has much room for improving its understanding of the domains of ToM, the findings in this and other studies do not seem to point to ToM merely being a byproduct of general cognitive ability or intelligence.

While these findings suggest that ToM is better represented as a multidomain construct that presents individual differences across and within adults, it is possible that the model defined in this study varies in other populations. Specifically, much of the research conducted with children in ToM suggests that ToM develops by age 5 (see Wellman et al., 2001), however, research has shown that ToM does continue to develop and improve throughout childhood and adolescence when age-appropriate tasks are used (Miller, 2010) and adults present differences in ToM task performance as well. This developmental trend is in line with the development of general cognitive abilities and executive functions, which tend to be first detected in early childhood but improve until adulthood (Best & Miller, 2010) after which individual differences are commonly studied (e.g., Miyake et al., 2000). Thus, while different domains of EF tend to develop with slightly different trajectories, they overall form a unitary construct. In fact, the relationships among some tasks in children tend to be similar to those in adults (e.g., Warnell & Redcay, 2019), however this might be due to the tasks used and the differences in ToM domains. Overall, it is possible that ToM follows a similar trend. Future research should examine how domains of ToM relate to each other in early age, as well as developmental changes in the construct.

This study focuses on performance among neurotypical adults and it is possible that some of the findings vary in other populations commonly studied in ToM research, such as clinical populations and children. However, as mentioned above, there are reasons to belief that some of the findings are in line with research studying the relationship between ToM and EF more broadly, perhaps indicating that the models tested in this study could be similar in other populations. Indeed Osterhaus et al. (2016) also found that a multidomain solution fit the data better than a single solution in a group of children after controlling for multiple cognitive abilities. In addition, studying ToM in adults is also informative of a skill that was largely considered "full-fledged" in adults (Keysar et al., 2003), but that has since been reconceptualized as a changing ability that varies along a continuum. Examining the degree to which individual differences in ToM influence other health and life outcomes should be considered as a way to provide support for the predictive validity of the construct. Further studying this ability could help understand whether improved ToM has beneficial outcomes that are worth fostering in many aspects of daily life, such as schools and workplaces.

It is worth noting that the relationship between RMET and Raven's might be partly driven by task characteristics, in addition to shared cognitive processes. For example, both tasks have a strong visuospatial and perceptual domain that requires matching items with potential responses and identifying and remembering rules. However, it could be argued that the director task also shares some of these properties but presents a weaker relationship with Raven's. Interestingly, the RMET also presented strong correlation with Gc (Table 2) as reported in the literature (Baker et al., 2014; Olderbak et al., 2015), but weaker partial correlations in the NMA (Fig. 3), and the edges in the network all seem pass largely through Gf nodes. This suggests that perhaps the relationships between the RMET and Gc are based largely on task characteristics, while the relationship between the RMET and Gf is not only based on similarities with superficial stimuli. The nature of these relationships should be further investigated.

Similarly, the SSQ may present inherent task demands that contribute to differences among tasks, above and beyond the processes

tapped by the tasks. The SSQ overall requires more reading comprehension skills and has no visual domain, which might be why there were notable differences with the other tasks in the network model. However, if the characteristics of the SSQ are responsible for these relationships, we would also observe stronger links with the Gc tasks that require more general reading and knowledge comprehension. However, this was not the case in the network model. Finally, the weak relationship between the RMET and the director task could also be related to aspects of the specificity of the tasks. However, considering that both tasks loaded into the same factor in the EFA and that their relationship seemed to depend on the relationship that both have with Raveńs based on the NMA, it is more likely that the fluid domain of the tasks is responsible for most of their similarities, and therefore when that process is controlled for, the relationship weakens. Overall, these suggests that most of the tasks had some degree of commonality, but their strong cognitive dependency accounts for much of this commonality. In future research, this could be avoided by selecting two or more tasks of each ToM domain studied so that the variance among the tasks is not affected so strongly by domain-general cognitive ability.

Recently, the director task has been the subject of intense debate. Specifically, researchers have questioned whether it measures a specific dimension of ToM or rather some other cognitive processes, such as mental rotation or selective attention (Rubio-Fernández, 2017). In this study, the director task was poorly related to the other ToM tasks but moderately correlated to the Gf and Gc tasks. It is unclear from this study whether the director task constitutes an adequate measure of ToM. It is possible that the director task represents a perspective-taking (cognitive), rather than social-cognitive or social-perceptual dimension of ToM, thus correlating more strongly with fluid reasoning. Another possibility is that the director task does not necessarily measure a ToM-related ability but a different cognitive process. For example, Rubio-Fernández (2017) has proposed that perhaps the director task could instead be measuring selective attention rather than solely ToM. Specifically, she proposes that the egocentric eye fixations often observed when participants see the critical items in the director task might not necessarily represent participants' egocentric fixations, but rather the use of selective attention to discard the inappropriate item that hinders the listener's ability to carry out an action. This question is still unclear and requires further experimental research, however the findings in this research open the possibility for the director task to be capturing some dimension of ToM while also requiring a degree of attention (largely related to fluid intelligence).

The RMET has also been the subject of criticism. Specifically, Oakley et al. (2016) suggested that using RMET as an affective task of ToM does not adequately captures pure affective mental states, but rather the ability to recognize complex emotional states. While this study did not differentiate between specific domains of ToM, the findings that alexithymia (i.e., lack of emotion recognition) predicts RMET has prompted researchers to suggest that more psychometric research is needed to specifically capture affective ToM that is separate from facial emotion recognition. However, these criticisms assume that facial emotion recognition can indeed be separate from affective ToM when it is also possible that it is part of a domain of ToM. In fact, researchers have proposed that it is possible that emotion recognition could be conceptualized as a mental state inference, based on theoretical frameworks that consider ToM a mix of processes that include social and cognitive abilities to infer mental states (Altschuler et al., 2021; Apperly, 2012a, Apperly, 2012b; Mitchell, 2005; Schaafsma et al., 2015; Wellman, 2014).

Finally, given the diverse nature of the ToM tasks, it is possible for domain-specific characteristics and variability in task demands to have contributed to the low correlations found among the ToM tasks. In this case, the low correlations might point to a larger role of domain-specific traits, as opposed to domain-general, among the ToM tasks as opposed to Gf and Gc tasks, potentially masking relationships to a general ToM construct. Future research should examine if tasks that share domain-specific traits load into a latent factor more strongly that tasks from different domains. This could indicate a need to assess ToM domains using multiple tasks depending on the goal of the researcher.

In general, these findings indicate that there are systematic issues around the conceptualization of ToM. Shaafsma et al. (2015) have pointed out that the use of ToM has been "vague and inconsistent" across the literature and that this provokes deep inconsistences in the research, both related to the validity of the measures and the claims made via the use of these measures in experimental designs. As discussed in the introduction, there are a number of different levels in which ToM is conceptualized and described, and different terminology is used to refer to the same construct, including cognitive development, social cognition, self-understanding, perception of others, mentalizing, understanding logical inferences, emotion and/or empathy. However, it is unlikely that all these different aspects of cognition correspond to the same umbrella term. As an example, in intelligence research, it is often thought that general intelligence encompasses a number of specialized abilities whose positive correlations (i.e., positive manifold) represent a general intelligence construct. For each of these abilities, a number of reliable and valid tests (ideally at least three) are necessary to measure each construct's validly and reliably. It is possible that ToM follows a similar structure, whereby multiple related but separate abilities or domains form theory of mind. Quesque and Rossetti (2020) have proposed that there are likely several separate mechanisms under the term ToM. For example, the RMET might assess "Facial Expression Categorization", the SSQ could be a task of "Mental States Ascription", and so on. To understand whether ToM is the result of multiple processes, it is necessary that multiple ToM measures reflect both general and specific computational processes used when one interprets desires, intentions, and beliefs in social contexts.

6. Conclusion

This study shows that the monolithic view of ToM is not supported by the processes measured by ToM tasks. To further explore the multiplicity of domains that form ToM, researchers should consider theoretical accounts of ToM that consider the dynamic nature of this ability and employ reliable and valid tests of *each* proposed domain to more adequately capture ToM ability. Conducting this important psychometric work would allow researchers to reconcile developmental, clinical, cognitive and neurological research on ToM.

Funding

This research was funded by Claremont Graduate University's 2020 Dissertation Award.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

OSF link included in manuscript

Acknowledgements

I would like to thank Andrew Conway, Kathy Pezdek, Megan Zirnstein, and Eleonora Rossi for their help in the revision of this manuscript.

References

Abu-Akel, A., & Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia, 49*(11), 2971–2984.
Altschuler, M. R., Trevisan, D. A., Wolf, J. M., Naples, A. J., Foss-Feig, J. H., Srihari, V. H., et al. (2021). Face perception predicts affective theory of mind in autism spectrum disorder but not schizophrenia or typical development. *Journal of Abnormal Psychology, 130*(4), 413.
Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology, 65*(5), 825–839.
Apperly, I. A., & Butterfill, S. (2009). Do Humans Have Two Systems to Track Beliefs and Belief-Like States? *Psychological Review, 116*(4), 953–970.
Avis, J., & Harris, P. L. (2016). Belief-Desire Reasoning among Baka Children : Evidence for a Universal Conception of Mind. *Society for Research in Child Development, 62*(3), 460–467.
Baker, C. A., Peterson, E., Pulos, S., & Kirkland, R. A. (2014). Eyes and IQ: A meta-analysis of the relationship between intelligence and "Reading the Mind in the Eyes". *Intelligence, 44*(1), 78–92. https://doi.org/10.1016/j.intell.2014.03.001
Baron-cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*, 37–46.
Baron-Cohen, S., O'Riordan, M., Jones, R., Stone, V., & Plaisted, K. (1999). A new test of social sensitivity: Detection of faux pas in normal children and children with Asperger syndrome. *Journal of Autism and Developmental Disorders, 29*(5), 407–418.
Baron-Cohen, S., Wheelwright, S., & Jolliffe, A. T. (1997). Is there a" language of the eyes"? Evidence from normal adults, and adults with autism or Asperger syndrome. *Visual cognition, 4*(3), 311–331.
Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241–251. https://doi.org/10.1017/S0021963001006643
Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V., & Lawson, J. (2001). Are intuitive physics and intuitive psychology independent? A test with children with Asperger Syndrome. *Learning, 5*(January 2014), 47–78. https://doi.org/10.1111/j.1469-7610.2004.00232.x
Behne, T., Carpenter, M., & Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental science, 8*(6), 492–499.
Bernhardt, B. C., & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience, 35*, 1–23.
Bernstein, D. M., Thornton, W. L., & Sommerville, J. A. (2011). Theory of mind through the ages: Older and middle-aged adults exhibit more errors than do younger adults on a continuous false belief task. *Experimental Aging Research, 37*(5), 481–502. https://doi.org/10.1080/0361073X.2011.619466
Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child development, 81*(6), 1641–1660.
Białecka-Pikul, M., Szpak, M., Zubek, J., Stępień-Nycz, M., Kołodziejczyk, A., Bosacki, S., et al. (2021). Measuring advanced theory of mind: Do story-based tasks work? *Journal of adolescence, 93*, 28–39.
Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*(1), B25–B31.
Bowman, L. C., & Wellman, H. M. (2014). Neuroscience contributions to childhood theory-of-mind development. *Contemporary perspectives on research in theories of mind in early childhood education, 2014*, 195–224.
Bowman, L. C., Liu, D., Meltzoff, A. N., & Wellman, H. M. (2012). Neural correlates of belief-and desire-reasoning in 7-and 8-year-old children: An event-related potential study. *Developmental Science, 15*(5), 618–632.
Brandone, A. C., & Wellman, H. M. (2009). You can't always get what you want: Infants understand failed goal-directed actions. *Psychological science, 20*(1), 85–91.
Brewer, N., Young, R. L., & Barnett, E. (2017). Measuring theory of mind in adults with autism spectrum disorder. *Journal of autism and developmental disorders, 47*(7), 1927–1941.
Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental Neuropsychology, 28*(2), 595–616.
Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development: An International Journal of Research and Practice, 11*(2), 73–92.
Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology, 87*(4), 299–319.
Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*(4), 1032–1053.
Carlson, S. M., Koenig, M. A., & Harms, M. B. (2013). Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(4), 391–402.
Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22.
Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Houghton Mifflin.
Chen, K. W., Lee, S. C., Chiang, H. Y., Syu, Y. C., Yu, X. X., & Hsieh, C. L. (2017). Psychometric properties of three measures assessing advanced theory of mind: Evidence from people with schizophrenia. *Psychiatry Research, 257*, 490–496.
Coyle, T. R., Elpers, K. E., Gonzalez, M. C., Freeman, J., & Baggio, J. A. (2018). General intelligence (g), ACT scores, and theory of mind:(ACT) g predicts limited variance among theory of mind tests. *Intelligence, 71*, 85–91.
Cutting, A. L., & Dunn, J. (2002). The cost of understanding other people: Social cognition predicts young children's sensitivity to criticism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 43*(7), 849–860. https://doi.org/10.1111/1469-7610.t01-1-00047
Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113–126.

de Abreu, P. M. E., Conway, A. R., & Gathercole, S. E. (2010). Working memory and fluid intelligence in young children. *Intelligence, 38*(6), 552–561.
De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science, 7*(1), 28–38.
Devaine, M., Hollard, G., & Daunizeau, J. (2014). Theory of mind: Did evolution fool us? *PloS One, 9*(2), e87619.
Devine, R. T., & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child Development, 84*(3), 989–1003.
Devine, R. T., & Hughes, C. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the silent films and strange stories tasks. *Journal of Experimental Child Psychology, 149*, 23–40.
Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: A new task for assessing theory of mind in adults. *PloS One, 8*(11).
Dumontheil, I., Apperly, I. A., & Blakemore, S. J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science, 13*(2), 331–338. https://doi.org/10.1111/j.1467-7687.2009.00888.x
Duncan, John, Johnson, Roger, Swales, Michaela, & Freer, Charles (1997). Frontal Lobe Deficits after Head Injury: Unity and Diversity of Function. *Cognitive Neuropsychology, 14*(5), 713–741. https://doi.org/10.1080/026432997381420
Eisenmajer, R., & Prior, M. (1991). Cognitive linguistic correlates of 'theory of mind' ability in autistic children. *British Journal of Developmental Psychology, 9*(2), 351–364. https://doi.org/10.1111/j.2044-835x.1991.tb00882.x
Ekstrom, R. B., Dermen, D., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests* (Vol. 102). Princeton, NJ: Educational testing service.
Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*(4), 617.
Epskamp, S., Lunansky, G., Tio, P., & Borsboom, D. (2018). Recent developments on the performance of graphical LASSO networks [Blog post].
Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.
Flavell, J. (1974). *The development of inferences about others*. In T. Mischel (Ed.), Understanding other persons. Rowman and Littlefield.
Flavell, J. (1977). *The development of knowledge about visual perception*. In Nebraska Symposium on Motivation, 43–76.
Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1-Level 2 distinction. *Developmental Psychology, 17*(1), 99–103. https://doi.org/10.1037/0012-1649.17.1.99
Frith, C. D. (2004). Schizophrenia and theory of mind. *Psychological Medicine, 34*(3), 385–389. https://doi.org/10.1017/S0033291703001326
Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1431), 459–473.
Frith, U., & Happé, F. (1994). Language and communication in autistic disorders. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 346*(1315), 97–104.
Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in cognitive sciences, 7*(2), 77–83.
Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin, 134*(1), 31.
German, T. P., & Hehman, J. A. (2006). Representational and executive selection resources in "theory of mind": Evidence from compromised belief-desire reasoning in old age. *Cognition, 101*(1), 129–152. https://doi.org/10.1016/j.cognition.2005.05.007
Gerrans, P., & Stone, V. E. (2008). Generous or parsimonious cognitive architecture? Cognitive neuroscience and theory of mind. *British Journal for the Philosophy of Science, 59*(2), 121–141. https://doi.org/10.1093/bjps/axm038
Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 312–7 years old on a stroop-like day-night test. *Cognition, 53*(2), 129–153.
Gibson, S., & Ninness, B. (2005). Robust maximum-likelihood estimation of multivariable dynamic systems. *Automatica, 41*(10), 1667–1682.
Gopnik, A., & Wellman, H. (1994). The 'theory' theory. In L. C. John Tooby, Alan M. Leslie, Dan Sperber, Alfonso Caramazza, Argye E. Hillis, Elwyn C. Leek, Michele Miozzo (Ed.), Mapping the Mind: Domain Specificity in Cognition and Culture. Cambridge: Press, Cambridge University.
Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085.
Gourlay, C., Collin, P., Caron, P. O., D'Auteuil, C., & Scherzer, P. B. (2020). Psychometric assessment of social cognitive tasks. *Applied Neuropsychology: Adult*, 1–19.
Greenberg, A., Bellana, B., & Bialystok, E. (2013). Perspective-taking ability in bilingual children: Extending advantages in executive control to spatial reasoning. *Cognitive Development, 28*(1), 41–50. https://doi.org/10.1016/j.cogdev.2012.10.002
Grice, H. (1989). *Studies in the Way of Words*. Harvard University Press.
Hala, S., Hug, S., & Henderson, A. (2003). Executive function and false-belief understanding in preschool children: Two tasks are harder than one. *Journal of Cognition and Development, 4*(3), 275–298.
Hamel, R., & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the Raven Advanced Progressive Matrices Test. *Educational and Psychological measurement, 66*(6), 1039–1046.
Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24*(2), 129–154. https://doi.org/10.1007/BF02172093
Harris, P. L. (2006). *Social Cognition*. In D. K. and R. S. W. Damon, R.M. Lerner (Ed.), Handbook of Child Psychology. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470147658.chpsy0219.
Hayward, E. O., & Homer, B. D. (2017). Reliability and validity of advanced theory-of-mind measures in middle childhood and adolescence. *British Journal of Developmental Psychology, 35*(3), 454–462.
Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science, 9*, 131–143.
Horn, J. L. (1994). Theory of fluid and crystallized intelligence. *Encyclopedia of Human Intelligence, 1*, 443–451.
Hughes, C., & Leekam, S. (2004). What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development. *Social Development, 13*(4), 590–619. https://doi.org/10.1111/j.1467-9507.2004.00285.x
Hughes, C., & Russell, J. (1993). Autistic children's difficulty with mental disengagement from an object: Its implications for theories of autism. *Developmental Psychology*. https://doi.org/10.1037//0012-1649.29.3.498
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Kan, K. J., Kievit, R. A., Dolan, C., & van der Maas, H. (2011). On the interpretation of the CHC factor Gc. *Intelligence, 39*(5), 292–302.
Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry, 27*(3), 151–177.
Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review, 9*(4), 637–671.
Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology, 26*, 165.
Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25–41. https://doi.org/10.1016/S0010-0277(03)00064-7
Kline, R. B. (2015). Principles and practice of structural equation modeling. *Guilford*. publications.
Lawson, J., Baron-Cohen, S., & Wheelwright, S. (2004). Empathising and systemising in adults with and without Asperger Syndrome. *Journal of Autism and Developmental Disorders, 34*, 301–310.
Lee, K., Olson, D. R., & Torrance, N. (1999). Chinese children's understanding of false beliefs: The role of language. *Journal of Child Language, 26*(1), 1–21. https://doi.org/10.1017/S0305000998003626
Leekam, S. R., & Prior, M. (1994). Can Autistic Children Distinguish Lies from Jokes? A Second Look at Second-order Belief Attribution. *Journal of Child Psychology and Psychiatry, 35*(5), 901–915. https://doi.org/10.1111/j.1469-7610.1994.tb02301.x
Legg, E. W., Olivier, L., Samuel, S., Lurz, R., & Clayton, N. S. (2017). Error rate on the director's task is influenced by the need to take another's perspective but not the type of perspective. *Royal Society Open Science, 4*(8), 170284.
Leslie, A. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. C. John Tooby, Alan M. Leslie, Dan Sperber, Alfonso Caramazza, Argye E. Hillis, Elwyn C. Leek, Michele Miozzo (Ed.), Mapping the Mind: Domain Specificity in Cognition and Culture (119–148).

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science, 1*(2), 247–253. https://doi.org/10.1111/1467-7687.00038
Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology, 50*(1), 45–85.
Masangkay, Z. S., McCluskey, K. A., McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The Early Development of Inferences about the Visual Percepts of Others. *Child Development, 45*(2), 357–366.
McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence., 37*(1), 1–10.
Miller, S. A. (2010). Social-Cognitive Development in Early Childhood. *Encyclopedia on Early Childhood Development.*
Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x
Mitchell, J. (2005). The false dichotomy between simulation and theorytheory: The argument's error. *Trends in Cognitive Sciences, 9*(8), 363–364. https://doi.org/10.1016/j.tics.2005.06.010
Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100.
Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology, 24*(3), 603–613.
Naito, M., Komatsu, S., & Fuke, T. (1994). Normal and autistic children's understanding of their own and others' false belief: A study from Japan. *British Journal of Developmental Psychology, 12*(3), 403–416. https://doi.org/10.1111/j.2044-835x.1994.tb00643.x
Navarro, E., & Conway, A. R. (2021). Adult bilinguals outperform monolinguals in theory of mind. *Quarterly Journal of Experimental Psychology, 74*(11), 1841–1851.
Navarro, E., Goring, S. A., & Conway, A. R. (2021). The relationship between theory of mind and intelligence: A formative g approach. *Journal of Intelligence, 9*(1), 11.
Nunnally, J. C., Jr (1970). *Introduction to psychological measurement.* McGraw-Hill.
Oakley, B. F., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of Abnormal Psychology, 125*, 818–823.
Obhi, S. (2012). The amazing capacity to read intentions from movement kinematics. Frontiers in Human Neuroscience, 6, Article 162. doi:10.3389/fnhum.2012.00162.
Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology, 6*, 1503.
Osterhaus, C., Koerber, S., & Sodian, B. (2016). Scaling of advanced theory-of-mind tasks. *Child Development, 87*(6), 1971–1991.
Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*, 34–72.
Perner, J., Lang, B., & Kloo, D. (2002). Theory of mind and self-control: More than a common problem of inhibition. *Child Development, 73*(3), 752–767.
Piaget, J., & Inhelder, B. (1956). *The child's concept of space.* Routledge & Paul.
Poletti, M., et al. (2012). Cognitive and affective Theory of Mind in neurodegenerative diseases: Neuropsychological, neuroanatomical and neurochemical levels. *Neurosci. Biobehav. Rev., 36*, 2147–2164.
Premack, D., & Woodruff, G. (1978). Chimpanzee theory of mind. *Behavioral and Brain Sciences, 4*(1978), 515–526.
Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences, 25*(1), 1–20.
Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science, 15*(2), 384–396.
Raven, J. C. (1938). Progressive Matrices: Sets A, B, C, D, and E. University Press, published by HK Lewis.
Rosi, A., Cavallini, E., Bottiroli, S., Bianco, F., & Lecce, S. (2016). Promoting theory of mind in older adults: Does age play a role? *Aging & Mental Health, 20*(1), 22–28.
Rubio-Fernández, P. (2017). The director task: A test of Theory-of-Mind use or selective attention? *Psychonomic Bulletin & Review, 24*(4), 1121–1128.
Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1255.
Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87–124.
Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences, 19*(2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007
Scholl, B. J., & Leslie, A. M. (2001). Minds, modules, and meta-analysis. *Child development, 72*(3), 696–701.
Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews, 42*, 9–34.
Sebanz, N., & Shiffrar, M. (2009). Detecting deception in a bluffing body: The role of expertise. *Psychonomic Bulletin & Review, 16*(1), 170–175.
Sodian, B. (1991). The development of deception in young children. *British Journal of Developmental Psychology, 9*(1), 173–188. https://doi.org/10.1111/j.2044-835x.1991.tb00869.x
Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *American Journal of Psychology, 15*, 201–293.
Spearman, C. E. (1927). *The abilities of man* (Vol. 89). New York: Macmillan.
Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.
Tager-Flusberg, H., & Sullivan, K. (2000). A domainial view of theory of mind: Evidence from Williams syndrome. *Cognition, 76*(1), 59–90.
Tardif, T., & Wellman, H. M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology, 36*(1), 25–43.
Thurstone, L. L. (1938). *Primary mental abilities* (Vol. 119). Chicago: University of Chicago Press.
Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71*, 1–26.
Van der Meer, L., Groenewold, N. A., Nolen, W. A., Pijnenborg, M., & Aleman, A. (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying Theory of Mind. *NeuroImage, 56*(4), 2364–2374.
Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *Neuroimage, 48*(3), 564–584.
van Veluw, S. J., & Chance, S. A. (2014). Differentiating between self and others: An ALE meta-analysis of fMRI studies of self-recognition and theory of mind. *Brain Imaging and Behavior, 8*(1), 24–38.
Wang, Z., & Su, Y. (2013). Age-related differences in the performance of theory of mind in older adults: A dissociation of cognitive and affective domains. *Psychology and Aging, 28*(1), 284.
Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition, 191*, 103997.
Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.
Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology, 15*(6), 728–755.
Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523–541.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684. https://doi.org/10.1111/1467-8624.00304
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraircing function of wrong belifs in young children's understanding of deception. *Cognition, 13*, 103–128.
Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913–934.

Further reading

Gopnik, A., Meltzoff, A., & Kuhl, P. (2000). *The scientist in the crib: What early learning tells us about the mind.* William Morrow Paperbacks.
Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*(1), 79–132.
Happé, F., & Frith, U. (1996). Theory of mind and social impairment in children with conduct disorder. *British Journal of Developmental Psychology, 14*(4), 385–398. https://doi.org/10.1111/j.2044-835x.1996.tb00713.x
Rubio-Fernández, P., & Glucksberg, S. (2012). Reasoning about other people's beliefs: Bilinguals have an advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(1), 211.